Information Networks

Hacettepe University
Department of Information Management
DOK 422: Information Networks

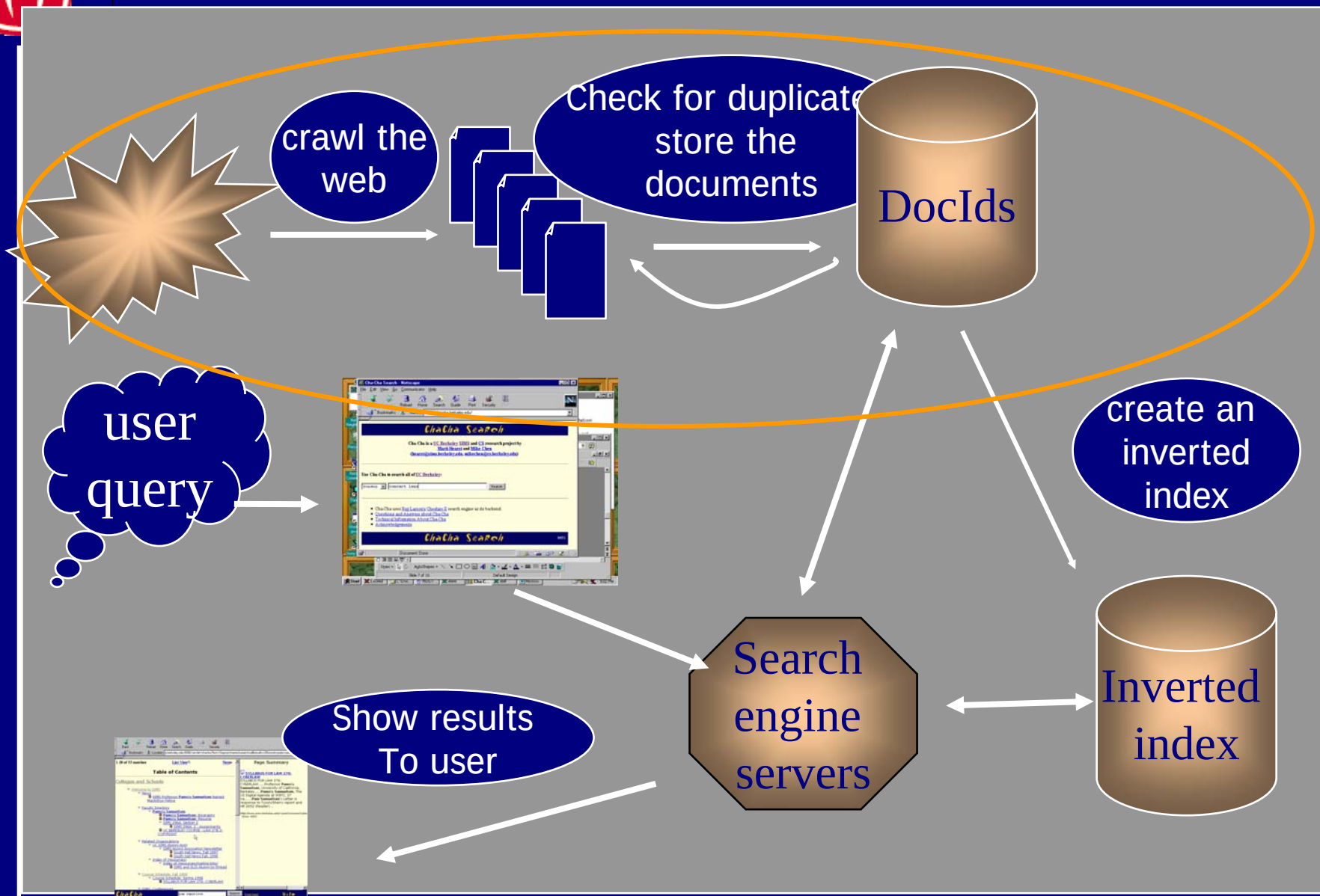
Search engines

Some Slides taken from: Ray Larson

Search engines

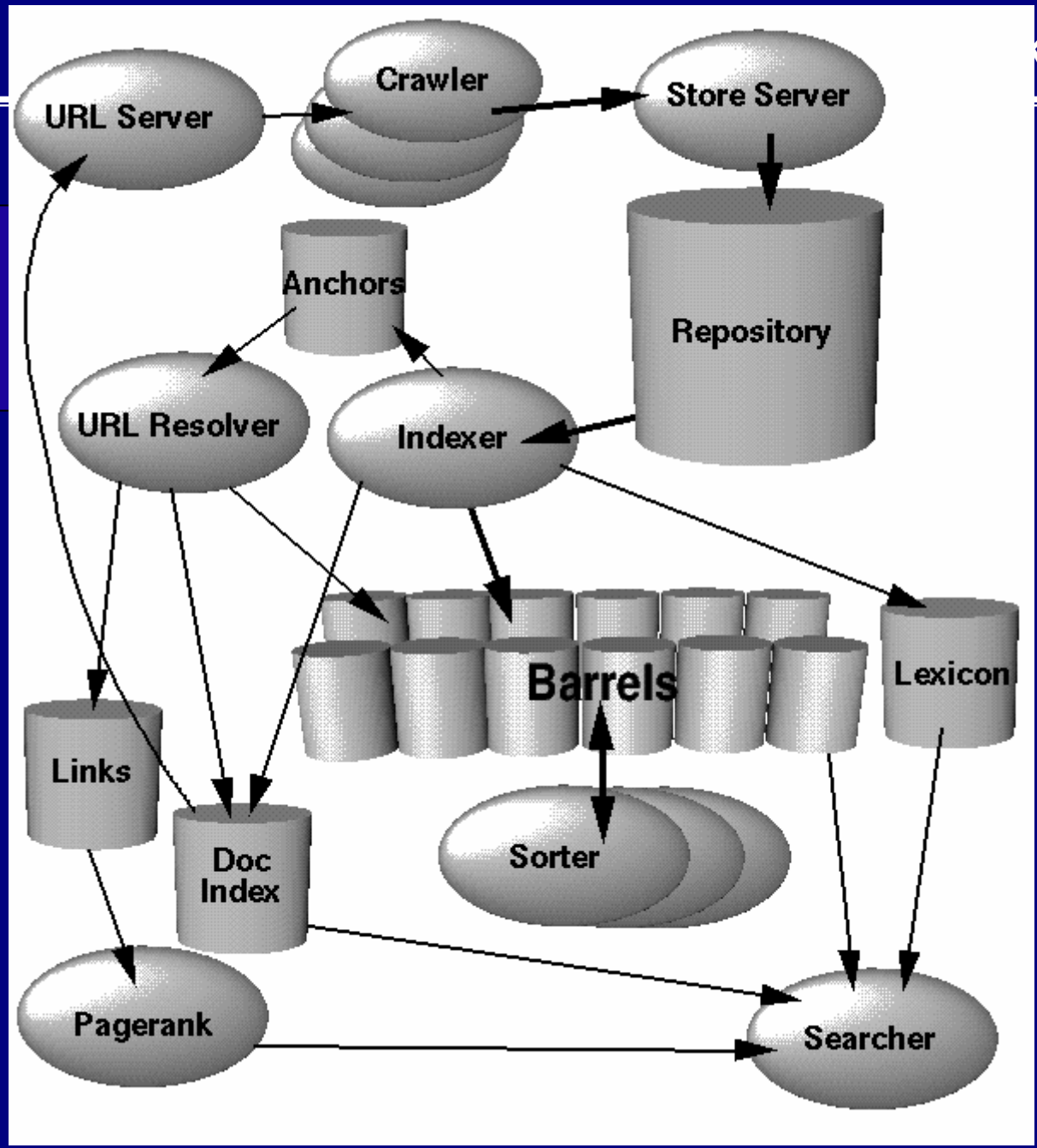
- Web Crawling
- Web Search Engines and Algorithms

Standard Web Search Engine



More detailed architecture, from Brin & Page 98.

Only covers the preprocessing in detail, not the query serving.



Web Crawling

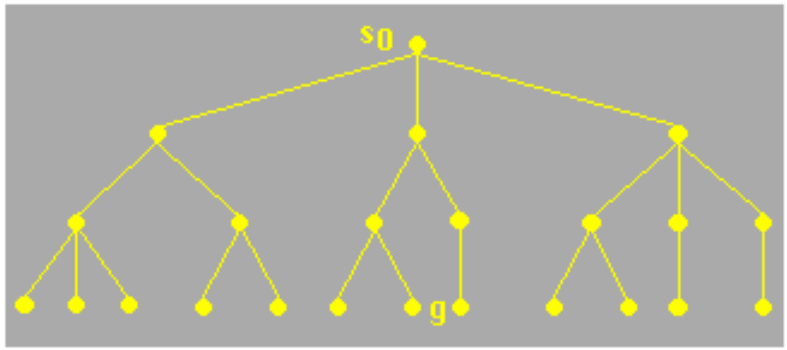
- How do the web search engines get all of the items they index?
- Main idea:
  - Start with known sites
  - Record information for these sites
  - Follow the links from each site
  - Record information found at new sites
  - Repeat

Web Crawlers

- How do the web search engines get all of the items they index?
- More precisely:
  - Put a set of known sites on a queue
  - Repeat the following until the queue is empty:
    - Take the first page off of the queue
    - If this page has not yet been processed:
      - Record the information found on this page
        - Positions of words, links going out, etc
      - Add each link on the current page to the queue
      - Record that this page has been processed
- In what order should the links be followed?

Page Visit Order

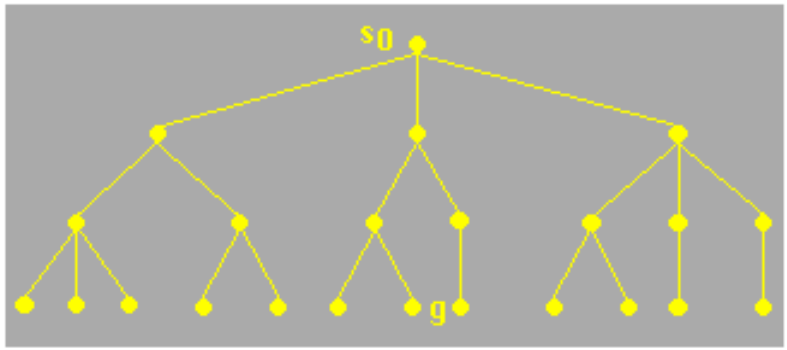
☞ http://www.rci.rutgers.edu/~cfs/472_html/AI_SEARCH/ExhaustiveSearch.html



Structure to be traversed

Page Visit Order

http://www.rci.rutgers.edu/~cfs/472_html/AI_SEARCH/ExhaustiveSearch.html



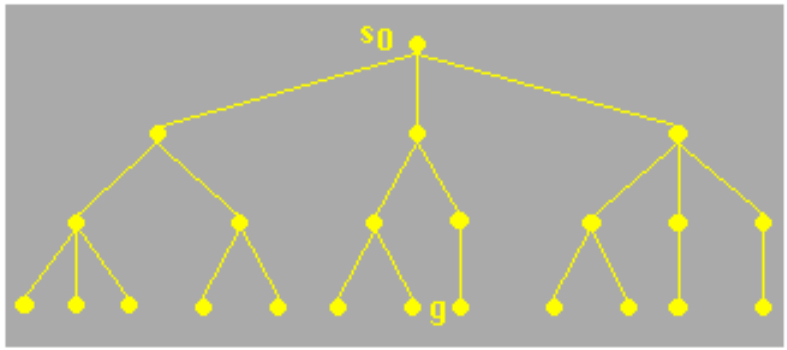
Breadth-first search

(must be in presentation mode to see this animation)

Page Visit Order

☞

http://www.rci.rutgers.edu/~cfs/472_html/AI_SEARCH/ExhaustiveSearch.html

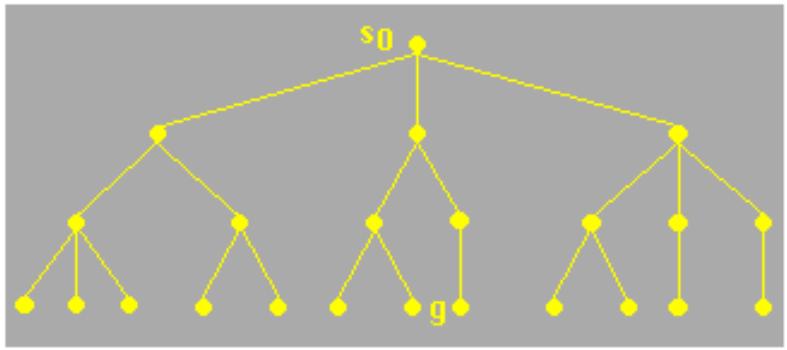


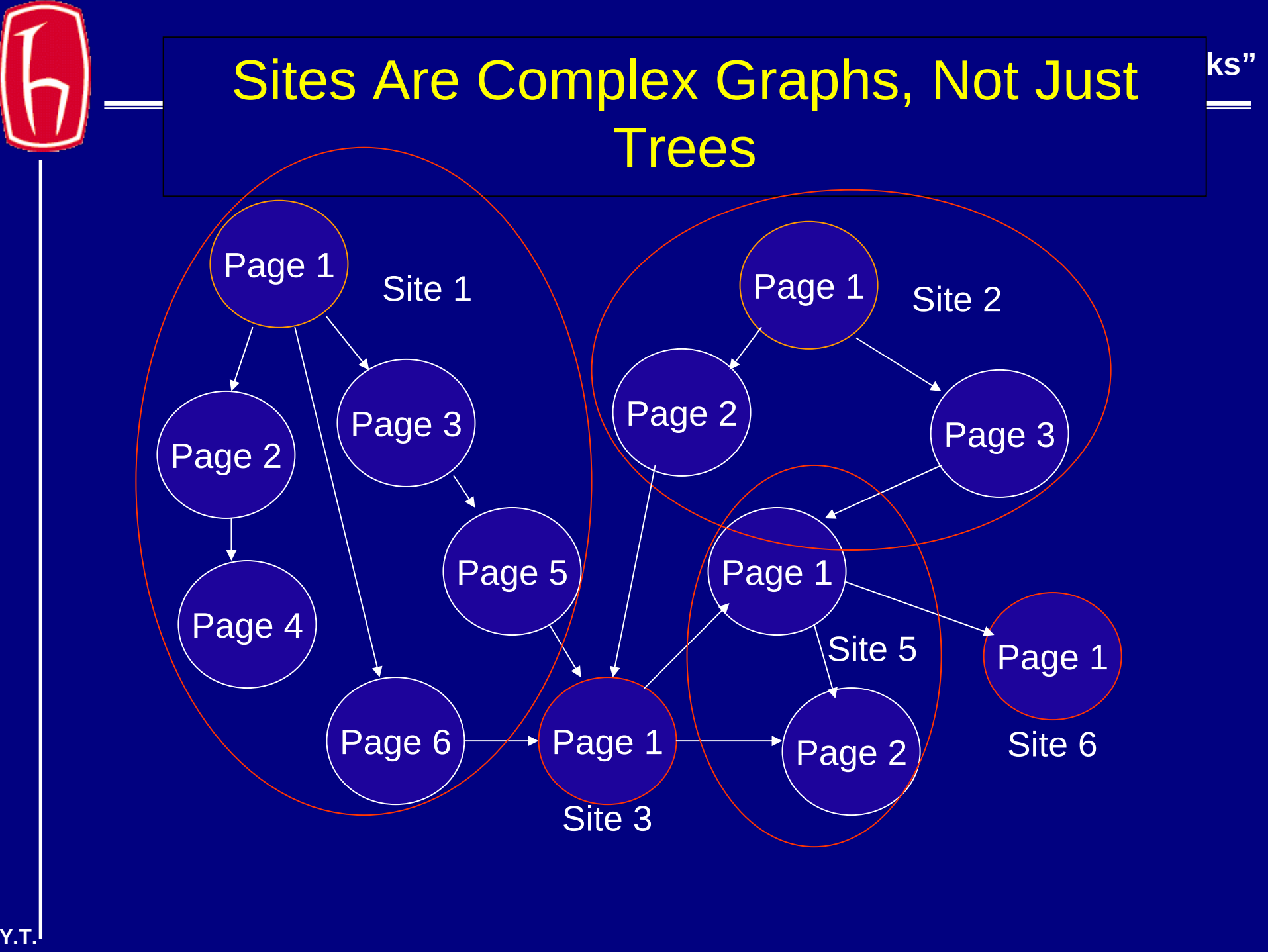
Depth-first search

(must be in presentation mode to see this animation)

Page Visit Order

http://www.rci.rutgers.edu/~cfs/472_html/AI_SEARCH/ExhaustiveSearch.html




Sites Are Complex Graphs, Not Just Trees
Page 1
Site 1
Page 2
Page 3
Page 4
Page 5
Page 6
Page 1
Site 3
Page 1
Site 2
Page 2
Page 3
Page 1
Site 5
Page 2
Page 1
Site 6

Web Crawling Issues

- Keep out signs
  - A file called robots.txt tells the crawler which directories are off limits
- Freshness
  - Figure out which pages change often
  - Recrawl these often
- Duplicates, virtual hosts, etc
  - Convert page contents with a hash function
  - Compare new pages to the hash table
- Lots of problems
  - Server unavailable
  - Incorrect html
  - Missing links
  - Infinite loops
- Web crawling is difficult to do robustly!

Searching the Web

- Web Directories versus Search Engines
- Some statistics about Web searching
- Challenges for Web Searching
- Search Engines
  - Crawling
  - Indexing
  - Querying

Directories vs. Search Engines

- Directories
  - Hand-selected sites
  - Search over the contents of the descriptions of the pages
  - Organized in advance into categories

- Search Engines
  - All pages in all sites
  - Search over the contents of the pages themselves
  - Organized after the query by relevance rankings or other scores

Search Engines vs. Internal Engines

- Not long ago HotBot, GoTo, Yahoo and Microsoft were all powered by Inktomi
- Today Google is the search engine behind many other search services (such as Yahoo up until very recently and AOL’s search service)

Statistics from Inktomi

- Statistics from Inktomi, August 2000, for one client, one week
  - Total # queries: 1315040
  - Number of repeated queries: 771085
  - Number of queries with repeated words: 12301
  - Average words/ query: 2.39
  - Query type: All words: 0.3036; Any words: 0.6886; Some words:0.0078
  - Boolean: 0.0015 (0.9777 AND / 0.0252 OR / 0.0054 NOT)
  - Phrase searches: 0.198
  - URL searches: 0.066
  - URL searches w/http: 0.000
  - email searches: 0.001
  - Wildcards: 0.0011 (0.7042 '?'s)
    - frac '?' at end of query: 0.6753
    - interrogatives when '?' at end: 0.8456
    - composed of:
      - who: 0.0783 what: 0.2835 when: 0.0139 why: 0.0052 how: 0.2174 where 0.1826 where-MIS 0.0000 can,etc.: 0.0139 do(es)/did: 0.0

What Do People Search for on the Web?

- Topics
  - Genealogy/Public Figure: 12%
  - Computer related: 12%
  - Business: 12%
  - Entertainment: 8%
  - Medical: 8%
  - Politics & Government 7%
  - News 7%
  - Hobbies 6%
  - General info/surfing 6%
  - Science 6%
  - Travel 5%
  - Arts/education/shopping/images 14%

(from Spink et al. 98 study)

Challenges for Web Searching: Data

- Distributed data
- Volatile data/”Freshness”: 40% of the web changes every month
- Exponential growth
- Unstructured and redundant data: 30% of web pages are near duplicates
- Unedited data
- Multiple formats
- Commercial biases
- Hidden data

Challenges for Web Searching: Users

- Users unfamiliar with search engine interfaces (e.g., Does the query “apples oranges” mean the same thing on all of the search engines?)
- Users unfamiliar with the logical view of the data (e.g., Is a search for “Oranges” the same things as a search for “oranges”?)
- Many different kinds of users

Web Search Queries

- Web search queries are SHORT
  - ~2.4 words on average (Aug 2000)
  - Has increased, was 1.7 (~1997)
- User Expectations
  - Many say “the first item shown should be what I want to see”!
  - This works if the user has the most popular/common notion in mind